CAE

A Case Study of an International Performance-Based Assessment of Critical Thinking Skills

Raffaela Wolf
Doris Zahner
Fiorella Kostoris
Roger Benjamin

Introduction

The measurement of higher-order competencies within a tertiary education system across countries presents methodological challenges due to differences in educational systems, socio-economic factors, and perceptions as to which constructs should be assessed (Blömeke, Zlatkin-Troitschanskaia, Kuhn, & Fege, 2013). According to Hart Research Associates (2009), there is substantial merit in assessing twenty-first century skills such as critical thinking and writing since about 78% of academic institutions in the United States have established cross-discipline learning outcomes, so called meta domains (Porter, McMaken, Hwang, & Yang, 2011), that all undergraduate students should possess upon graduation. Furthermore, changing skill demands of graduating students have been observed around the world since the 1990s (Levy & Murname, 2004). Meeting the demands of today's world requires a shift in assessment strategies to measure the skills now prized in a complex global environment. More specifically, assessments that only foster the recall of factual knowledge have been on the decline, whereas assessments that evoke higher-order cognitive skills have seen an accelerating demand in the twenty-first century. As an example, CAE (the Council for Aid to Education) has been developing assessments that target higher-order skills. The Collegiate Learning Assessment-plus (CLA+) is a measure that emulates critical-thinking and writing skills.

In late 2012, the Agenzia Nazionale di Valutazione del Sistema Universitario e della Ricerca (ANVUR) approached CAE proposing a research study to test the feasibility of adapting, translating, and administering CLA+ to higher education students in Italy. The purpose of this feasibility study was twofold. The first purpose was to see if it was possible to assess Italian students' higher-order skills as outlined in Table 1. The second purpose was to see if the Italian students' performance was comparable to their American counterparts.

It is evident that these types of competencies are desirable in many cultures around the globe, regardless of discipline or curriculum. However, measuring competencies within an international framework poses psychometric challenges that pertain to test development, scoring, and the validity of score interpretations (Hambleton & Murphy, 1992). Bias and measurement equivalence (ME) are two different, yet intertwined, pivotal notions that pertain to instrument characteristics in cross-cultural comparisons. Bias is often referred to as nuisance, or confounding factors, whereas equivalence is related to issues concerning the measurement of the instrument (Van de Vijver, 1998). Different forms of bias are considered the main sources of in-equivalence in cross-cultural research (Van de Vijver, 1998; Van de Vijver & Leung, 1997). Bias occurs when observed results systematically distort the relationships between true scores and observed variables. Thus, bias is considered a threat to the validity of the score inferences drawn within a cross-cultural context. There are two main forms of bias: construct and method, where the former refers to unintended differences in the latent constructs, while the latter represents differences in the process of measurement that are due to characteristics of the instrument or administration. Item bias was not considered in the current study.

Construct comparability rests upon the assumption that test scores are contingent upon the same definition of higher-order skills across the countries. If the constructs are comparable, then test score differences across countries may reflect a true representation of the discrepancies in student performance. However, within the context of such comparisons, differences in scores may be influenced by confounding variables, such as test adaptation (e.g. translation), familiarity with item response formats, and many other socio-cultural factors, which introduce method bias. For example, selected-response items (SRQs) are widely used in the United States, whereas many European countries make use of performance or constructed-response tasks (Wolf, 1998). The lack of familiarity with a particular item type could create a source of construct irrelevant variance and, thus, limit the validity of score interpretations. A mixed-format type assessment, consisting of both performance tasks (PTs) and SRQs, can be deemed a viable option in an attempt to ensure test fairness and to reduce the potential impact of bias across cultures.

CLA+ is a mixed-format type assessment; thus this paper presents the results from the feasibility study as a case study of the successful adaption, translation, and administration of CLA+ in 12 Italian institutions. A discussion is provided regarding how different biases may be addressed within an international context. A second analysis examined whether students from Italy and the US ascribe the same meanings to different item formats (PT and SRQs) thus addressing the issue of measurement equivalence and the feasibility of cross-cultural score comparisons. Results are interpreted within a validity framework.

Methodology

Task Selection, Translation, and Adaption of CLA+

CLA+ consists of two sections, a PT and a set of SRQs. ANVUR was presented with an assortment of PT and SRQ sets and a committee of bilingual educators and administrators decided upon the "Parks" PT and a set of SRQs that they felt were culturally appropriate and adaptable for use in the Italian context. The PT and SRQs were then translated and adapted by a third party translation group and eventually verified by ANVUR and CAE staff. ANVUR was provided with a translation and adaptation guide to help facilitate the process. Following the translation and adaptation of the PT and SRQs, ANVUR conducted cognitive labs and a small pilot study, with Italian university students, to verify that the translated and adapted version of CLA+ was clear and elicited the appropriate types of student responses.

CAE adapted its current CLA+ Testing Platform ("CLA+ Platform") to accommodate the adaptation and translation changes made to the "Parks" PT and the 25 SRQs. CAE implemented an additional platform, encompassing text translations as necessary, to facilitate the administration of the tests in Italy. The CLA+ Platform was modified to accommodate student responses in Italian.

Participants

ANVUR recruited 12 universities to participate in this feasibility study, four from three geographical regions (i.e., north, central, and south). The student participants from the 12 universities (n = 5853) comprised of graduating students in their third and fourth year at their respective institutions. These students took the Italian CLA+ during the spring semester of 2013. A sample of American students (n = 4666) were selected for comparative purposes. The American student participants were university freshmen from the fall semester of 2013. The sampled institutions (public and private) consisted of small liberal arts colleges, as well as large research institutions, from the various regions of the United States. Because CLA+ is a newly modified and upgraded version of CLA, the only comparison group available for this study was entering freshmen.

Test Administration

The Italian CLA+ was administered on ANVUR's testing platform. Students had a total of 90 minutes to complete the CLA+, 60 minutes for the PT, and 30 minutes for 20 SRQs. The American students had a similar administration of CLA+ except through a different test delivery platform. The test administration of the Italian CLA+ was vetted and approved by CAE, prior to administration, to assess comparability of the testing platforms. A customized testing platform was created for the Italian students so that testing conditions were uniform between the two countries.

CLA+

CLA+ is a performance-based authentic measure that targets higher-order competencies, such as critical-thinking and written-communication skills, by using a combination of both PTs and SRQs. The adapted version of the CLA+ consisted of one PT and 20 SRQs. Higher-order skills are emulated by presenting authentic tasks, within real-world contexts, in which students must demonstrate those skills. The PTs are designed so that students must get to the bottom of a problem and recommend a course of action after analyzing a document library that contains various sources of information, such as letters, maps, and graphs, just to name a few. As shown in Table 1, the PT is composed of three subscales: analysis and problem solving (identifying, interpreting, evaluating, and synthesizing pertinent information and proposing a solution in terms of how to proceed in case of uncertainty), writing effectiveness (producing an organized and cohesive essay with supporting arguments), and writing mechanics (demonstrating command of standard written English). Similarly to the PT, the SRQs are also developed with the intent to elicit higher-order cognitive skills rather than the recall of factual knowledge. Students are presented with a set of questions that pertain to documents from a range of information sources. The SRQ subscales were identified as critical reading and evaluation

(eight items), scientific and quantitative reasoning (seven items), and critique an argument (five items). Students were given 60 minutes to construct a response to the PT and 30 minutes to respond to the 20 SRQs.

Table 1
CLA+ Tasks and Subscales

Task	Subscales
PT	Analysis and Problem Solving
	Writing Effectiveness
	Writing Mechanics
SRQ	Critical Reading and Evaluation
	Scientific and Quantitative Reasoning
	Critique an Argument

Scoring

The PT of the adapted version of CLA+ was scored in Italy by a team of trained scorers. CAE representatives led a series of trainings both virtually and on-site in Rome. All responses were assigned raw subscale scores and raw total scores that reflected critical-thinking and writing skills. Total CLA+ scores were computed as a weighted sum of the PT (weighted at .50) and SRQs (weighted at .50).

For the PTs, CAE measurement scientists initially trained three scorers from ANVUR via Skype, followed by an additional in-person training of the Italian lead scorers (one representative from each participating institution plus the three scorers from ANVUR) in Rome. The ANVUR scorers prepared a translated version of the CAE scoring rubric. This team of Italian lead scorers then trained a set of Italian scorers to complete the scoring of the student PT responses.
The CLA+ scoring rubric for the PTs consists of three subscores: Analysis and Problem Solving (APS), Writing Effectiveness (WE), and Writing Mechanics (WM). Each of these subscales is scored from a range of 1–6, where 1 is the lowest level of performance and 6 is the highest, with each score pertaining to specific response attributes. For all task types, blank or entirely off-topic responses are flagged for removal from results. Because each prompt may have differing possible arguments or relevant information, scorers receive prompt-specific guidance in addition to the scoring rubrics. Additionally, the reported subscores are not adjusted for difficulty like the overall CLA+ scale scores, and, therefore, are not directly comparable to each other. These PT subscores are intended to facilitate criterion-referenced interpretations, as defined by the rubric.

Analysis and Problem Solving (APS) measures a student's ability to make a logical decision or conclusion (or take a position) and support it with accurate and relevant information (facts, ideas, computed values, or salient features) from the document library.
Writing Effectiveness (WE) assesses a student's ability to construct and organize logically cohesive arguments. This is accomplished by strengthening the writer's position by elaborating on facts or ideas (e.g., explaining how evidence bears on the problem, providing examples, and emphasizing especially convincing evidence).

Writing Mechanics (WM) evaluates a student's facility with the conventions of standard written English (agreement, tense, capitalization, punctuation, and spelling) and control of the English language, including syntax (sentence structure) and diction (word choice and usage).
The selected-response section of CLA+ consists of 20 items distributed across three subscales: scientific and quantitative reasoning (seven items), critical reading and evaluation (eight items), and critique an argument (five items). Subscores in these sections are determined according to the number of questions correctly answered, with scores adjusted for the difficulty of the particular question set received.

Data Analysis

Independent sample t-tests were conducted to assess whether there were significant mean differences on the PT and SRQs across countries. In an attempt to examine whether students accredit the same meaning to the different item formats, a multi-group confirmatory factor analysis (MG-CFA) was conducted (Byrne, Shavelson, & Muthén, 1989). In the first step, a confirmatory factor analysis (CFA) model was specified that reflected how higher-order skills

were theoretically operationalized. A one-factor CFA model, a two-factor CFA model and a higher-order CFA model were tested. The two-factor model had the best model fit in both countries:

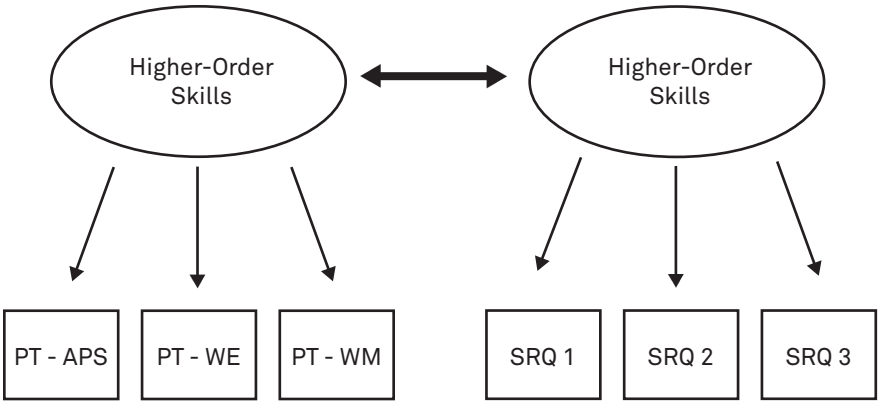

Figure 1. Example of Correlated Traits Model with 3 PT subscales and 3 SRQs

This model was fitted for the American and Italian students separately to ensure that the same model is valid in each group. Secondly, a baseline model was established by running a common model for both groups with unconstrained parameters. In the third step, several models were estimated to test for ME:

Table 1
Testing for Measurement Invariance with Categorical Data

Model	Factor loadings	Thresholds	Residual variances	Factor means	Factor Variances
Configural invariance	*	*	Fixed at 1	Fixed at 0	Fixed at 1
Strong invariance (1)	Fixed	Fixed	Fixed at 1	Fixed at 0/*	Fixed at 1
Strong invariance (2)	Fixed	Fixed	Fixed at 1	Fixed at 0/*	Fixed at 1/*

*Note. The * indicates that the parameter is freely estimated. Fixed at 0/*= the factor means are fixed at 0 in one group and freely estimated in the other group. Fixed at 1/* = the factor variance is fixed at 1 in one group and freely estimated in the other group.*

The various models were fit using an adjusted weighted least squares (WLSM) algorithm using the Mplus software (Muthén & Muthén, 2010). All model in this analysis were evaluated in terms of goodness of fit criteria. Exact fit was evaluated using the model $\chi 2$, whereas close fit was evaluated using the comparative fit index (CFI), Tucker-Lewis non-normed fit index (TLI), and root mean squared error of approximation (RMSEA). In this study, values of less than .05 were used for the RMSEA and values greater than .95 were used for the TLI (Hu & Bentler, 1999). All fit indices were used conjunctively to determine model fit.

Results

Descriptive Statistics

Table 1 provides descriptive statistics for the adapted CLA+. Both countries showed similar results for the PT (Italy: M = 9.17, SD = 2.95 ; US: M = 9.06, SD = 2.54), whereas the sample from Italy had a higher mean on the SRQs (M = 12.31, SD = 2.85) compared to the American sample (M = 10.64, SD = 3.62). Independent sample t-tests showed statistically significant differences on the SRQs (t (10564) = 25.82, p<.001) but not on the PT task. However, it is uncertain whether these differences are due to true differences in performance or whether the familiarity with item types across cultures introduced nuisance variability.

Table 1
Descriptive statistics for CLA+ for Italian vs. American students

	Italy		US	
	SRQ	PT	SRQ	PT
Items (N)	20	1	20	1
Students (N)	5853	5853	4638	4638
Min Score	0	3	0	3
Max Score	19	18	19	18
Mean	12.31	9.17	10.64	9.06
SD	2.85	2.95	3.62	2.54

Factor Analyses Results

The first step was to test whether the proposed two-factor model fits the empirical data for each group. Results indicate that the hypothesized model is supported in both groups (Italian: $\chi 2 = 1280.05$; df = 229; RMSEA = .028; CFI = .989; TLI = .988; American: $\chi 2 = 2203.51$; df = 229; RMSEA = .043; CFI = .992; TLI = .992). The second step was to move from a single-group CFA to MG-CFA in order to cross-validate the two-factor model across the two groups (configural invariance). Table 1 indicates that Model 1 provided a good fit ($\chi 2 = 3455.13$; df = 458 ; RMSEA = .035 ; CFI = .99; TLI = .99) to the data, indicating that the factorial structure of the construct is equal across the two groups. In other words, examinees ascribe the same meaning to the definition of higher-order skills across countries. Given that configural invariance was confirmed, the factor loadings and thresholds were then constrained to be equal to test for strong invariance. Model 2 fit significantly worse than Model 1, DIFFTEST(56) = 13239.55, $p<.001$, and Model 3 fit significantly worse than Model 2, DIFFTEST(2) = 1402.13, $p<.001$. These results suggest that students may have ascribed different meanings to the item formats across countries.

Table 2
Fit indices for invariance tests

	χ^2	df	RMSEA	CFI	TLI
Model 1: Baseline (Configural invariance)	3455.13	458	.035	.99	.99
Model 2: Strong Invariance (1)	25166.68	514	.095	.92	.92
Model 3: Strong Invariance (2)	23764.55	512	.093	.92	.92

Table 3
Fit indices for model comparisons

	χ^2	p
Model 1vs Model 2	13239.55	<.001
Model 2 vs Model 3	1402.13	<.001

Discussion

The feasibility of assessing higher-order skills in two different cultures was confirmed in this study. Cross-cultural studies aim to address the question to whether valid test score inferences can be drawn across different cultural populations. This case study was an attempt to address bias as a function of the interpretation of test scores rather than an inherent property of the instrument. It is well known that test adaptations or translations are prone to introducing different types of biases (Hambleton, 1996), such as construct, method, and item bias. In this feasibility study, translation effects were mitigated through the implementation of a multi-stage translation process. Through the combined effort of colleagues and content-area experts from each culture it was possible to specify and examine the similarities in the underlying construct definition of higher-order skills and the alignment of the items with the test blueprint. As part of the adaptation phase, a small pilot study was conducted in Italy to ensure that items on the instrument were functioning as intended. Consequently, it was determined that Italian and

American students appear to associate the same meaning to the definition of higher-order skills and that the items on the instrument were adequately sampled from the domain of higher-order skills. The appropriateness of construct representativeness across countries was confirmed by the results of the CFA analyses.

Method bias may be introduced through administration procedures and/or differences that pertain to the instrument itself. The test administration platform of the Italian CLA+ was examined by CAE prior to administration to ensure comparability of the testing platforms. In order to circumvent problems due to rater effects, specific scoring rubrics and guidelines were developed, and graders underwent rigorous training sessions that were facilitated through the joint effort of both countries. However, there was reason to believe that the use of different item formats could be a source of method bias since familiarity with item types varies by culture (Wolf, 1998). Post-hoc statistical analyses were conducted in an attempt to examine whether examinees from Italy and the US ascribe the same meaning to the PT and SRQs. According to these results, it is evident that higher-order skills were assessed in both countries. However, students appeared to associate different meanings with different item types across countries, which imposes the question as to whether valid score inference can be drawn from direct score comparisons of students in different countries. Psychometric evidence exists for providing valid score inferences within each country due to the successful adaptation of CLA+. However, direct score comparisons across countries should be made with caution because a total score that is comprised of PT and SRQ scores may have an altered meaning in both countries due to the dissimilar meanings that are associated with different item types. This could be due to the differences in the two populations, which is a limitation of the current study. CLA+ is a newly modified and upgraded version of the CLA; thus, the only comparison group available for this study was entering freshmen who were compared to graduating students in Italy. This implies that the groups may have varied in ability, which was not accounted for in the analyses. Plans for a future analysis include the use of U.S. CLA+ senior data in order to examine whether the effect of growth in higher-order skills from freshmen to graduating seniors may have had an impact on the results of the current study. Furthermore, when interpreting the test scores across countries, other factors that could impact test score results, such as student motivation and/or socio-economic status, need to be addressed.

During the last few decades, bias has predominantly been associated with item bias or differential item functioning; methods to address construct and method bias often appear to be neglected. While the importance of addressing item bias is evident in cross-cultural research, it is also apparent that cross-cultural comparisons can further be challenged by construct irrelevant sources of variance that go beyond individual items. Perhaps an ongoing effort, including both a priori and post-hoc considerations, could provide fruitful information in terms of construct, method, and item bias. Rather than viewing and/or treating each component in isolation, a holistic approach that combines these sources could ensure high standards in all stages of the test development and adaptation process, consequently aiding in the collection of evidence for valid cross-cultural score interpretations.

Some suggestions for future a priori activities include a focus on collaborative efforts between measurement scientists, cognitive scientists, and experts within the tertiary education system from both cultures in an attempt to develop instruments that are within appropriate cultural contexts. Different translation procedures also may be combined to ensure adequate translations. The translated instrument could be pilot tested with bilingual students to assess the appropriateness of the adapted version. However, findings may need to be interpreted with caution since the bilingual students may not be representative of the target population. In an attempt to minimize method bias, it may be worthwhile to provide practice items so that students from different cultures can become accustomed to different item formats. Individual items also should be reviewed in terms of poor translation, complex wording of items, and whether items invoke unintended additional abilities. Statistical analyses at the item level, such as differential item functioning, should be integrated into the item development process to ensure appropriateness of translated items. Comparisons of item statistics in the two versions of the instrument should consider controlling for any ability differences in the two groups.

Bias is often perceived as a nuisance factor (Van de Vijver, 1998) and thus many statistical procedures exist in an attempt to mitigate or reduce the unwanted effects of bias on cross-cultural score comparisons. However, if bias would be neglected, then perhaps one could gain information in terms of systematic cross-cultural differences, which may indeed be beneficial to the instrument development process. This would also aid in the collection of validity evidence to ensure appropriate cross-cultural comparisons. In sum, it is feasible to assess higher-order skills globally. However, in a collaborative effort across nations, numerous factors need to be taken into consideration prior, during, and after the test adaptation phase to ensure that valid cross-cultural score inferences can be drawn from the data.

References

Arum, R. & Roksa, J. (2011). *Academically Adrift: Limited Learning on College Campuses*. Chicago, Ill.: University of Chicago Press.

Blömeke, S., Zlatkin-Troitschanskaia, O., Kuhn, C., & Fege, J. (2013). *Modeling and Measuring Competencies in Higher Education: Springer*.

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). *Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance.* Psychological Bulletin, 105(3), 456.

Hambleton, R. K. (1996). *Guidelines for Adapting Educational and Psychological Tests*.

Hambleton, R. K., & Murphy, E. (1992). *A psychometric perspective on authentic measurement*. Applied Measurement in Education, 5(1), 1-16.

Hart Research Associates. (2009). *Learning and Assessment: Trends in Undergraduate Education - A Survey Among Members of The Association of American Colleges and Universities*. Washington, DC: Hart Research Associates.

Hu, L. t., & Bentler, P. M. (1999). *Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives*. Structural Equation Modeling: A Multidisciplinary Journal, 6(1), 1-55.

Levy, F., & Murname, R. J. (2004). *Education and the Changing Job Market: An Education Centered on Complex Thinking and Communicating is a Graduate's Passport to Prosperity*. Educational Leadership, 62(2), 80-83.

Muthén, B., & Muthén, L. (2010). Mplus Version 6.1 [Software]. Los Angeles, CA: Author.

Porter, A., McMaken, J., Hwang, J., & Yang, R. (2011). *Common Core Standards: The New US Intended Curriculum*. Educational Researcher, 40(3), 103-116.

Van de Vijver, F. J. (1998). *Towards a theory of bias and equivalence*. Zuma Nachrichten, 3, 41-65.

Van de Vijver, F. J., & Leung, K. (1997). *Methods and data analysis for cross-cultural research (Vol. 1): Sage*.

Wolf, R. M. (1998). *Validity issues in international assessments*. International journal of educational research, 29(6), 491-501.